벡터와 행렬의 미적분

1. 스칼라를 벡터로 미분하는 경우

함수의 출력변수가 스칼라이고 입력변수 $\mathbf{x}$가 벡터인 다변수 함수를 사용하는 경우 결과를 열벡터로 표시한다.

이렇게 만들어진 벡터를 Gradient Vector라고 하고 ∇f로 표기한다.

$$\nabla f = \frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

Gradient vector는 각 변수로의 일차 편미분 값으로 구성되는 벡터이고, 이 벡터는 f의 값이 가장 가파르게 증가하는 방향을 나타낸다.

또한 이 때 그 벡터의 크기는 그 증가하는 방향의 정도(= 기울기)를 의미한다.

반대로 gradient에 음수를 취하면, 즉 $\nabla - f$는 f가 가장 가파르게 감소한느 방향을 나타내게 된다.

이러한 negative gradient는 어떤 함수를 지역적으로 선형근사하거나 gradient descent 방법으로 함수의 극점을 찾는 용도로 사용된다.

Gradient를 이용한 다변수 스칼라 함수 f의 어떤 점 p에서의 선형 근사식은 테일러 근사를 이용해 다음과 같이 표현할 수 있다.

$$f(x) \simeq f(p) + \nabla f(p)(x - p)$$

2. 행렬미분법칙

I. 행렬미분법칙 1: 선형 모형

선형모형을 미분하면 gradient vector는 가중치 벡터가 된다.

$$f(x) = w^\top x$$

$$\nabla f = \frac{\partial w^\top x}{\partial x} = \frac{\partial x^\top w}{\partial x} = w$$

II. 행렬미분법칙 2: 이차형식

이차형식을 미분하면 행렬과 벡터의 곱으로 나타난다.

$$f(x) = x^\top Ax$$

$$\nabla f(x) = \frac{\partial x^\top Ax}{\partial x} = (A + A^\top)x$$

3. 벡터를 벡터로 미분하는 경우

벡터 $\mathbf{x}$를 입력받아 벡터를 출력하는 함수 $f(\mathbf{x})$를 생각하자.

벡터를 벡터로 미분하면 미분을 당하는 원소가 여러개이고 미분을 하는 벡터의 원소도 여러개이므로 미분의 결과인 도함수는 2차원 배열, 즉 행렬이 된다.

$$\frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial \mathbf{x}} & \frac{\partial f_2}{\partial \mathbf{x}} & \cdots & \frac{\partial f_N}{\partial \mathbf{x}} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_N}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_N}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_M} & \frac{\partial f_2}{\partial x_M} & \cdots & \frac{\partial f_N}{\partial x_M} \end{bmatrix}$$

I. 행렬미분법칙 3: 행렬과 벡터의 곱의 미분

행렬 A와 벡터 $\mathbf{x}$의 곱 $A\mathbf{x}$를 벡터 $\mathbf{x}$로 미분하면 행렬 $A^\top$가 된다.

$$f(x) = A\mathbf{x}$$

$$\nabla f(x) = \frac{\partial (A\mathbf{x})}{\partial \mathbf{x}} = A^\top$$

함수의 출력변수와 입력변수가 모두 벡터 데이터인 경우에는 입력변수 각각과 출력변수 각각의 조합에 대해 모든 미분이 존재한다.

따라서 도함수는 행렬 형태가 되며, 이렇게 만들어진 도함수의 행렬을 자코비안 행렬이라고 한다.

자코비안은 어떤 다변수 벡터함수에 대한 일차 미분의 형태라고 볼 수 있다.

그레디언트는 다변수 스칼라함수에 대한 일차 미분인 반면 자코비안은 다변수 벡터함수에 대한 일차 미분이라는 차이가 있다.

자코비안 행렬은 벡터함수를 벡터변수로 미분해서 생기는 행렬의 전치행렬이기 때문에 행과 열의 방향이 다르다는 점을 유의하자.

$$Jf(x) = \boldsymbol{J} = (\frac{\partial f}{\partial \mathbf{x}})^\top = \begin{bmatrix} (\frac{\partial f_1}{\partial \mathbf{x}})^\top \\ \vdots \\ (\frac{\partial f_M}{\partial \mathbf{x}})^\top \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_M}{\partial x_1} & \cdots & \frac{\partial f_M}{\partial x_N} \end{bmatrix}$$

다변수 함수의 2차 도함수는 gradient vector를 입력변수 벡터로 미분한 것으로 헤시안 행렬이라고 한다.

헤시한 행렬은 gradient vector의 자코비안 행렬의 전치 행렬로서 정의된다.

$$Hf(x) = \boldsymbol{H} = J(\nabla f(x))^\top = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_N \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_N^2} \end{bmatrix}$$

헤시안은 critical point의 종류를 판별하는 데 활용될 수 있다.

어떤 함수의 일차미분이 0이 되는 지점을 critical point라고 하는데 함수의 극점(극대, 극소), saddle point 등이 이에 해당한다.

어떤 함수를 최적화하기 위해 극점을 찾으려면 일단 일차 미분을 하여 gradient가 0이 되는 지점을 찾는다. 그런데 이렇게 찾은 지점이 극점인지 또는 saddle point인지 알 수 없다.

이 때 헤시안을 이용한 이차미분값으로 이를 판별한다.

어떤 함수의 critical point에서 계산한 헤시안 행렬의 모든 eigenvalue가 양수이면 해당지점에서 함수는 극소, 모든 eigenvalue가 음수이면 극대, eigenvalue가 음과 양을 모두 가지면 saddle point인 것으로 판단한다.

이는 헤시안 행렬의 고유벡터는 해당 함수의 곡률이 큰 방향벡터를 나타내고, 그 때 고유값이 해당 고유벡터의 방향으로의 함수의 곡률을 나타내기 때문에 가능한 판별이다.

cf) 자주 쓰이는 미분의 꼴

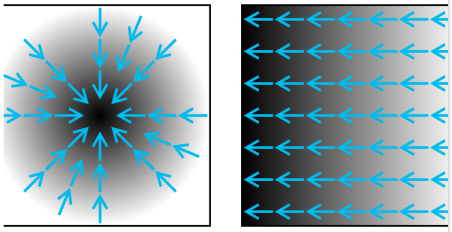
- $\mathbf{a} \cdot \mathbf{x} = \mathbf{a}^\top \mathbf{x} = \mathbf{x}^\top \mathbf{a} \quad \Rightarrow \quad \mathbf{a}^\top$
- $A\mathbf{x} \quad \Rightarrow \quad A$
- $\mathbf{x}^\top A \quad \Rightarrow \quad A^\top$
- $\mathbf{x}^\top A\mathbf{x} \quad \Rightarrow \quad \mathbf{x}^\top (A + A^\top)$
- $\mathbf{y}^\top \mathbf{z} = \mathbf{y} \cdot \mathbf{z} \quad \Rightarrow \quad \mathbf{y}^\top \frac{\partial \mathbf{z}}{\partial \mathbf{x}} + \mathbf{z}^\top \frac{\partial \mathbf{y}}{\partial \mathbf{x}}$
- $\|\mathbf{x}\|^2 = \mathbf{x} \cdot \mathbf{x} = \mathbf{x}^\top \mathbf{x} \quad \Rightarrow \quad 2\mathbf{x}^\top$
- $\|\mathbf{x}\| \quad \Rightarrow \quad \frac{\mathbf{x}^\top}{\|\mathbf{x}\|}$
- $\|A\mathbf{x} - \mathbf{b}\|^2 \quad \Rightarrow \quad 2(A\mathbf{x} - \mathbf{b})^\top A$차

▼ 참고

Gradient - Wikipedia

In vector calculus, the gradient of a scalar-valued differentiable function f of several variables is the vector field (or vector-valued function) whose value at a point is the

𝕎 https://en.wikipedia.org/wiki/Gradient



Jacobian matrix and determinant - Wikipedia

In vector calculus, the Jacobian matrix (,) of a vector-valued function of several variables is the matrix of all its first-order partial derivatives. When this matrix is square, that is, when the function takes the same number of variables as input as the number of vector components of its output, its determinant

𝕎 https://en.wikipedia.org/wiki/Jacobian_matrix_and_determinant

Hessian matrix - Wikipedia

In mathematics, the Hessian matrix or Hessian is a square matrix of second-order partial derivatives of a scalar-valued function, or scalar field. It describes the local curvature of a function of many variables. The Hessian matrix was developed in the 19th century by the German mathematician

𝕎 https://en.wikipedia.org/wiki/Hessian_matrix